Design of a competence-based assessment system for air traffic control training

Summary

Like other process control tasks in transportation (aviation, shipping, railways) or process industries (e.g., chemical and nuclear plants), the ATC task is considered a complex cognitive skill. Learnability seems to be limited. The strict safety requirements make the performance standards even higher because any error caused by incompetence of operators must be avoided. Consequently, the outcome from training is often too low despite of high selection standards. This may result in a shortage of competent personnel. Low pass rates are also undesirable because the training is usually very time-consuming and expensive. High-fidelity simulators are used to train trainees to the highest possible level before they enter the safety-critical working environment in on-the-job training (OJT).

A well-designed assessment system can contribute to solve this problem in several ways. First, assessment can support learning processes through feedback and adaptation of training to individuals' needs. This increases the chance for trainees to successfully complete the training. Second, reliable and valid pass-fail decisions can be produced by restricting the number of false positives, i.e. prospective-less trainees who unnecessarily occupy expensive training positions during later phases, and false negatives, i.e. failed trainees who could have had a chance to succeed training ultimately. Third, predictive validity of selection can be improved using more reliable training criteria in validation studies.

This study is the result of a project internally performed at Air Traffic Control the Netherlands (LVNL). Its objective was to design an assessment system that optimally supports the acquisition of air traffic controller competences in simulator and on-the-job training (OJT).

Design methodology

We combined two types of design methodologies, mainly based on the design cycle of Roozenburg and Eekels (1991) and the training design methodology of Moonen (2000). This resulted in a cyclic design process of three phases: (1) *analysis;* (2) *design and implementation*; (3) *evaluation*. The thesis is divided into three parts that correspond with these design phases. The first phase, *analysis,* comprises a literature review, an ATC competence analysis, and the construction of a Program of Requirements (PoR) with stakeholders. The second phase, *design and implementation*, refers to the design and development of the assessment system itself, and the introduction of the system in the training. The third phase, *evaluation*, involves the evaluation of the assessment system in three parts: (1) psychometric quality; (2)

learning processes (analysis of learning curves and competence development); (3) user evaluation. Finally, an overall evaluation against the Program of Requirements was held. We have learned that the design phases occur iteratively. A key success factor was the intensive collaboration with the user group (air traffic controllers) in designing an innovative product that was completely accepted for practical use.

Analysis

We started with a literature review that covers: (1) general theories about performance, competences and learning; (2) domain-specific literature about the nature of the ATC task and learning processes in ATC; (3) relevant literature about assessment that could be used in the context of ATC training. The findings were used in the design of the new assessment system.

Next, we did a competence analysis by organizing two workshops. Controllers were involved as subject matter experts (SME's) to explicate their implicit knowledge. Thirteen competences, represented by a set of behavioural markers, were formulated in their own jargon. This should enhance recognizability and common understanding for practical use. The next step was to operationalize the competences. We compared the list with published (cognitive) task analyses looking for additional aspects that might have been forgotten. This process resulted in the ATC Performance Model, which has served as a framework for the assessment design at LVNL (Oprins, Burggraaff & Van Weerdenburg, 2006).

Finally, a Program of Requirements (PoR), a set of requirements and constraints on the new assessment system, was constructed together with a group of stakeholders. We analysed the shortcomings of the previous system and the purposes of the redesign. The PoR has primarily guided the design of the new assessment system.

Design and implementation

The design process occurred in strong collaboration with the user group (managers, coaches and trainees) as part of the design teams. This has enhanced the implementation of the assessment system. We made choices in the design of the system to fulfill its purposes, and this has led to some specific properties.

The assessment system is competence-based. This implies that competences are assessed: the successful integration of knowledge, skills and attitudes and their application in realistic environments. Competences optimally reflect the individual abilities to perform effectively without paying attention to situational factors that may influence performance. All aspects that belong to competences are assessed (technical, cognitive, emotional, social) to get a complete picture. The competences were directly derived from the ATC Performance Model. This has resulted in the following set of competences for ACC: *safety, efficiency, verbal expression, listening, coordination, equipment operation, strip/label management, mental picture, attention management, planning, decisiveness, workload management, attitude,* and *teamwork ability.* Since they are not directly visible, each competence is represented by a set of *performance criteria*, formulated as 'behavioural markers' and rated at a six

points rating scale. During training, the same competences are assessed in order to follow the trainee's progression on each competence over a training period, based on various task situations and circumstances. The generic character of competences makes this possible, for instance, *planning* is relevant in any ATC task execution whether simple or complex. Progression provides an important indicator of whether a trainee is still learning or has reached a learning plateau. Potential deficiencies of trainees can be detected in an early stage so that training can be maximally adapted to the trainee's needs.

The competences are assessed against augmenting *performance standards* during training. Simulator training (preOJT) and OJT comprise successive phases with specific standards for each phase (cf. 'norms'). The standards were formulated as exemplary behaviours, a variant on behaviourally anchored rating scales (BARS). They do not specify scale positions but standards to be achieved at the end of each phase. In this way, assessors have more agreement in what is expected from trainees in intermediate phases. For trainees it is clearer which competences they have to develop further in a specific phase. The standards in preOJT are mainly defined by the sequence of simulator exercises. The classification into phases is based on two main variables: *traffic handling* (degree of safety and efficiency required) and *traffic complexity* (degree of task difficultness). Structuring OJT is more difficult due to the ongoing live traffic and was therefore most innovative. The division in phases in OJT, four phases in total, was based on three main principles: *traffic handling*, *traffic complexity*, and *aid of the coach*. The length of each phase is flexible, dependent on the trainee's progression

Continuous assessment is applied to get a complete picture of the trainee's performance and to measure progression over time. Coaches are continuously in interaction with trainees and can force them to verbalize their thoughts. Assessment of cognitive processes is required to obtain diagnostic information on performance and especially on performance shortcomings. *Progression reports* are filled in after one or two weeks of training in preOJT and OJT. The time intervals vary because of the different operational schedules. The coach has a double role as an assessor and a coach, and multiple assessors are involved. In preOJT, additional *simulator tests* are applied to measure trainees' performance objectively at a certain moment of time and in well-balanced test scenarios without interference by the coach. Pass-fail decisions are made at the end of each phase in preOJT and OJT. There does not exist a fixed cut-off, but decisions are based on an extensive quantitative and qualitative analysis of simulator tests and progression reports.

A web-based assessment tool is designed to fill in progression reports, to store the results in a database, and to generate several overviews of trainee performance. In this way, interested persons who have access from several places can better follow the trainee's progression over time. Training results can easily be used for reliability and validity studies. The first version was made in the assessment tool Questionmark Perception. A project has started to design a new tool because of some technical problems.

Much effort was put in the implementation of the assessment system, as this was considered very important for its correct use and acceptance. The close collabora-

tion with the users made that the system was accepted rather easily. Furthermore, we introduced the system carefully by starting with small pilots and by organizing presentations. We paid extra attention to assessment in coach and assessor training. Intermediate evaluations resulted in small changes before the final version was launched.

Evaluation

The requirements and constraints in the PoR guided the evaluation that comprised three parts. The data were from 34 trainees in ACC preOJT (188 progression reports; 36 simulator tests), and from 27 trainees in ACC OJT (407 progression reports), collected during the period January 2001 until December 2006.

Evaluation of psychometric quality
We investigated the following reliability issues: (1) *interrater agreement*; (2) *rating errors* (halo error, leniency and severity error, range restriction and central tendency error); (3) *test reliability* (internal consistency, split-half reliability). Predictive validity was part of the evaluation of learning processes (see next paragraph) because this involves analyses over time.

Interrater agreement was investigated by using simulator tests in which two assessors filled in test reports independently of each other. Their ratings should be interchangeable since assessments may not be dependent on the person who assesses. We examined three types of indices in interrater agreement: shape, dispersion, and level. The results have shown that the interrater agreement between assessors is sufficiently high for the overall performance level (weighted sum of competence ratings), but only moderate with respect to profile similarity (shape, dispersion). This implies that some assessors give low ratings on some competences, while others give high ratings on the same competences and vice versa. Apparently, assessors find it difficult to point at specific deficiencies of trainees.

Next, the presence of the most important rating errors was examined. Analyses of *leniency* and *severity errors* showed that assessors tend to leniency as expected: the mean competence ratings are above the scale midpoint. We found a few systematic differences in leniency between assessors: there exist some 'Santa Claus' (high means) and 'Axeman' (low means) assessors. Furthermore, assessors generally show *range restriction*. They do not equally distribute their ratings over the 6-points rating scale, but tend to give positive ratings and avoid the extreme scale points; thus, *central tendency* does not occur. Differences between assessors in their distribution of ratings, expressed in congruency indexes, were not found. Finally, we examined the occurrence of *halo errors*. We found high intercorrelations between some competence ratings. We recognized the classification of the ATC Performance Model in the results. Specific competences should be intercorrelated because they are conceptually related to each other. Therefore, we concluded that the presence of halo errors is not severe. In general, rating errors were found more often in progression reports. This confirms that simulator tests are more reliable.

Finally, we examined *internal consistency* and *split-half reliability*. Calculations of the Cronbach's alphas and item-total correlations per competence have shown

that the internal consistency is very high. Only a small number of criteria should be deleted or changed. Split-half reliability was estimated for simulator tests with the Spearman-Brown formula. We found very high reliability coefficients. Therefore, we considered test reliability of the assessment system as sufficient.

In sum, the findings have shown that the reliability of the assessment system is sufficient to a certain extent. The system is well designed in terms of the classification in competences, performance criteria and performance standards. However, the role of the assessor can be improved. Assessor training is organized to pay attention to the avoidance of rating errors and to achieve a common understanding of the competences.

Evaluation of learning processes

A well-designed assessment system works adequately if the assessment results represents learning processes adequately. Patterns and individual differences in learning (e.g., slow starters, learning plateaus) and in performance (strengths and weaknesses) should be clearly distinguished to provide a basis for adequate feedback and interventions. Learning curves can be derived from the assessment results based on a sequence of performance measures over time. These learning curves should be sufficiently representative for learning. Then the training can be maximally adaptive to the trainee's needs. Pass-fail decisions will be more valid if they are based on the predictability of patterns in learning processes.

Learning curves are usually presented as growth curves based on repeated executions of the same task at successive moments of time, but this is not possible with our assessment system. The trainee's performance is assessed against augmenting performance standards, which are constantly translated into the same 6-points rating scale. Therefore, the learning curves produced by our assessment system could rather be referred to as recalibrated learning curves. We compared the learning curves derived from progression reports in preOJT and OJT with prototypical learning curves that are expected from general learning theory. We defined three groups: (1) *high performers* (passed without problems); (2) *moderate performers* (passed with difficulties); (3) *low performers* (failed). Three training managers classified the trainees into the groups based on expert judgment, serving as an external criterion for trainee success.

We distinguished two quantitative variables that define the learning curves, subdivided into four measures: (1) *performance*: mean performance level; occurrences of insufficient performance; (2) *progression*: growth; rate of growth. We examined the distinction between the three groups visually and quantitatively. The results showed that the learning curves of actual trainees are sufficiently representative for learning processes, because they reflect the prototypical learning curves adequately. Although the number of trainees is too low yet for getting convincing evidence, we also explored the predictability of learning curves over time. The findings suggested that learning curves are predictive for next phases to a certain extent, but it is too early for defining strict cut-offs.

A well-designed competence-based assessment system should not only be able to represent general learning processes as expressed in the learning curves, but also the development of specific competences. This is important for identifying

possible deficiencies. We used the same classification into the three groups of performers. The findings were in conformance with the literature with respect to trainability of competences. As expected, we found significant differences between individual trainees for more critical and less trainable competences such as *mental picture* and *workload management*, than for competences that are less critical and more trainable such as *label management* and *equipment operation*. The main reasons for failing are respectively: *mental picture, workload management, attention management* and *decisiveness*. We concluded that the assessment system sufficiently represents individual differences in learning processes. More longitudinal research is needed to get evidence of predictive validity.

User evaluation
Multiple methods were used to evaluate the assessment system with the users (trainees, coaches, training managers). The findings have shown that the system is practically usable and has certainly improved learning processes. The tools, procedures and assessment reports are clear.

Conclusions and further research

The purpose of designing an assessment system for LVNL that makes training more efficient and effective was achieved. Learning processes are better supported and pass-fail decisions are more reliable and valid. The majority of the requirements and constraints in the PoR were fulfilled. Unfortunately, we do not have convincing evidence for an increased outcome of competent controllers from training yet.

More generally, we gained insight into assessment design for complex skill acquisition, based on the application in the ATC domain. However, a limitation of this thesis is that the findings are only based on a small sample. Some issues need further research. More research on the design of assessment systems for complex skill acquisition in simulator and on-the-job training should be done with special attention for the reliability and validity of human assessors' ratings. Furthermore, more insight into complex skill acquisition is needed to achieve that assessment can support learning optimally. Our new method of making learning curves may contribute to the field of modeling learning processes. In addition, trainability of competences in ATC and in related domains should be further examined. Its relationship with innate cognitive abilities and personality characteristics should be investigated in order to improve the predictive validity of selection and training. Moreover, further research is required on the relationship between adaptive training and deficiencies in competences. We experienced many difficulties with adapting training to the trainee's needs, while this seems to be a key success factor in completing the training. Possible interventions are: (dynamic) task selection, specific coaching, remedial teaching and counseling, alternative training trajectories, lengthening training, retraining, etcetera. Finally, advanced technology and automated measurements can help to improve assessment systems.

Samenvatting

Zoals andere proces control taken in transport (luchtvaart, scheepvaart, spoorwegen) of procesindustrie (bijv. chemische fabrieken en kerncentrales), wordt ook luchtverkeersleiding beschouwd als een complexe cognitieve taak. De leerbaarheid blijkt beperkt te zijn. De strenge eisen aan de veiligheid maken de vereiste prestaties nog hoger omdat elke fout, veroorzaakt door incompetentie van de operators, voorkomen moet worden. Ten gevolge hiervan is het aantal mensen dat de opleiding haalt vaak te laag, ondanks een strenge selectie. Dit kan leiden tot een tekort aan vakbewaam personeel. Lage slagingspercentages zijn bovendien onwenselijk omdat de opleiding doorgaans erg veel tijd en geld kost. Simulators van hoge kwaliteit worden gebruikt om leerlingen op te leiden tot het hoogst haalbare niveau, voordat ze de echte werkomgeving betreden waar de veiligheid kritisch is.

Een goed ontworpen beoordelingssysteem kan bijdragen tot de oplossing van dit probleem in meerdere opzichten. Ten eerste, beoordeling kan leerprocessen ondersteunen door middel van feedback en het aanpassen van de opleiding aan de behoeftes van individuele leerlingen. Dit verhoogt de kans dat de leerlingen de opleiding succesvol afronden. Ten tweede, zo'n beoordelingssysteem maakt het mogelijk om betrouwbare en valide pass-fail beslissingen te nemen. Het aantal 'false positives' moet zo beperkt mogelijk blijven, d.w.z. kansloze leerlingen die onnodig dure opleidingsplekken bezet houden tijdens latere fasen in de opleiding. Dit geldt ook voor het aantal 'false negatives', d.w.z. gezakte leerlingen die de opleiding uiteindelijk wel hadden kunnen halen. Ten derde, de predictieve validiteit van de selectie kan verbeterd worden door betrouwbare criteria uit de opleiding te gebruiken voor validatieonderzoek.

Deze studie is het resultaat van een project dat intern is uitgevoerd bij Luchtverkeersleiding Nederland (LVNL). Het doel was om een beoordelingssysteem te ontwerpen dat het aanleren van de competenties voor luchtverkeersleiding optimaal ondersteunt in de simulatoropleiding en in de werkplekopleiding.

Ontwerpmethodologie

We hebben twee ontwerpmethodologieën gecombineerd die hoofdzakelijk gebaseerd zijn op de ontwerpcyclus van Roozenburg en Eekels (1991) en de methodologie voor opleidingsontwerp van Moonen (2000). Dit heeft geleid tot een cyclisch ontwerpproces, bestaande uit drie fasen: (1) *analyse;* (2) *ontwerp en implementatie;* (3) *evaluatie.* De dissertatie is onderverdeeld in drie delen die overeenkomen met deze drie fasen. De eerste fase, *analyse*, bestaat uit literatuuronderzoek, een competen-

tieanalyse van luchtverkeersleiding, en het opstellen van een Programma van Eisen (PvE) samen met stakeholders. De tweede fase, *ontwerp en implementatie*, betreft het ontwerp en de ontwikkeling van het beoordelingssysteem zelf, en de introductie van het systeem in de opleiding. De derde fase, *evaluatie,* bevat de evaluatie van het beoordelingssysteem in drie delen: (1) psychometrische kwaliteit; (2) leerprocessen (analyse van leercurves en competentieontwikkeling); (3) gebruikersevaluatie. Tenslotte is een algehele evaluatie ten opzichte van het Programma van Eisen gehouden.

We hebben ondervonden dat de ontwerpfasen iteratief zijn. Een sleutel tot succes is de intensieve samenwerking met de gebruikers (luchtverkeersleiders) geweest in het ontwerpen van een innovatief product dat volledig geaccepteerd is voor toepassing in de praktijk.

Analyse

We zijn gestart met literatuuronderzoek, bestaande uit: (1) algemene theorieën over prestaties, competenties en leerprocessen; (2) domeinspecifieke literatuur over de aard van de taak van luchtverkeersleiders en over leerprocessen in luchtverkeersleiding; (3) relevante literatuur over beoordeling die toepasbaar is in de context van luchtverkeersleiding. De bevindingen zijn gebruikt in het ontwerp van het nieuwe beoordelingssysteem.

Vervolgens hebben we een competentieanalyse uitgevoerd door middel van het organiseren van twee workshops. Luchtverkeersleiders namen hieraan deel. Zij waren de inhoudsdeskundigen die hun impliciete kennis expliciet hebben gemaakt. Dertien competenties, elk ondersteund met een set van gedragscriteria, werden geformuleerd in hun eigen jargon. Dit zou de herkenbaarheid en algemeen begrip moeten stimuleren ten behoeve van het praktisch gebruik. De volgende stap was het operationaliseren van de competenties. We hebben de lijst van competenties met gepubliceerde (cognitieve) taakanalyses vergeleken om te onderzoeken of we wellicht aspecten vergeten waren. Dit proces leidde tot het zgn. 'ATC Performance Model'. Dit model heeft gediend als algemeen kader voor het ontwerp van het beoordelingssysteem bij LVNL (Oprins, Burggraaff & Van Weerdenburg, 2006a).

Tenslotte is er samen met een groep stakeholders een Programma van Eisen (PvE) opgesteld, bestaande uit een set van eisen en randvoorwaarden aan het nieuwe beoordelingssysteem. We hebben de tekortkomingen van het voorafgaande beoordelingssysteem geanalyseerd en de doelen voor het herontwerp afgeleid. Het PvE heeft richting gegeven aan het ontwerp van het nieuwe beoordelingssysteem.

Ontwerp en implementatie

Het ontwerpproces vond plaats in nauwe samenwerking met de gebruikers (managers, coaches en leerlingen) die onderdeel uitmaakten van de ontwerpteams. Dit heeft de implementatie van het systeem bevorderd. We hebben keuzes gemaakt in het ontwerp van het systeem om te kunnen voldoen aan de doelstellingen, en dit heeft geleid tot een aantal specifieke eigenschappen.

Het beoordelingssysteem is competentiegericht. Dit houdt in dat er wordt beoordeeld op competenties: de succesvolle integratie van kennis, vaardigheden en houdingen en hun toepassing in realistische omgevingen. Competenties hebben betrekking op de individuele capaciteiten om effectief te kunnen presteren, zonder dat er aandacht wordt besteed aan omgevingsfactoren die de prestaties kunnen beïnvloeden. Alle aspecten die behoren tot competenties worden beoordeeld (technisch, cognitief, emotioneel, sociaal) om een compleet beeld te krijgen. De competenties zijn direct afgeleid van het ATC Performance Model. Dit heeft geresulteerd in de volgende set van competenties voor ACC: *veiligheid, efficiëntie, uitdrukkingsvaardigheid, luistervaardigheid, coordinatie, omgaan met apparatuur, strip- en labelbehandeling, mentale beeldvorming, aandachtsverdeling, planning, besluitvaardigheid, omgaan met werkdruk, houding*, en *samenwerking*.

Omdat competenties niet direct zichtbaar zijn, wordt elke competentie ondersteund door een set van *gedragscriteria*, geformuleerd in waarneembaar gedrag, waarbij een zespunts ratingschaal wordt gehanteerd. Tijdens de opleiding wordt beoordeeld op dezelfde competenties om de voortgang van leerlingen op elke competentie te kunnen volgen gedurende een opleidingsperiode, gebaseerd op wisselende taaksituaties en omstandigheden. Het generieke karakter van competenties maakt dit mogelijk; zo is *planning* relevant in de uitvoering van elke taak bij luchtverkeersleiding, zowel bij een eenvoudige als complexe taak. *Progressie* is een belangrijke indicator of een leerling nog aan het leren is, of dat hij/zij een leerplateau heeft bereikt. Mogelijke problemen van leerlingen kunnen in een vroeg stadium worden ontdekt zodat de opleiding zo goed mogelijk kan worden aangepast aan de behoeftes van de leerling.

De competenties worden beoordeeld ten opzichte van oplopende *standaarden* tijdens de opleiding. De simulatoropleiding (preOJT) en werkplekopleiding (OJT) zijn opgebouwd uit fasen die elk specifieke standaarden hebben (vgl. 'normering'). De standaarden zijn geformuleerd in voorbeeldgedrag, een variant op de geankerde gedragsschalen (vgl. 'BARS'). Ze geven niet de schaalankers weer, maar het niveau dat bereikt moet zijn aan het eind van elke fase. Op deze manier komen de beoordelaars makkelijker tot overeenstemming over wat er van leerlingen wordt verwacht in tussenliggende fasen. Voor leerlingen is het duidelijker welke competenties zij verder moeten ontwikkelen in een bepaalde fase. De standaarden in de preOJT worden hoofdzakelijk bepaald door de opbouw van de simulatoroefeningen. De indeling in fasen is voornamelijk gebaseerd op twee variabelen: *verkeersafhandeling* (mate van veiligheid en efficiëntie) en *verkeerscomplexiteit* (mate van de moeilijkheid van de taak). Het structureren van de OJT is lastiger vanwege het voortdurende echte verkeer, en was daarom het meest innovatief. De indeling in fasen voor de OJT, vier fasen in totaal, is voornamelijk gebaseerd op drie principes: *verkeersafhandeling, verkeerscomplexiteit,* en *hulp door de coach*. De lengte van elke fase is flexibel en hangt af van de voortgang van de leerling.

Continuous assessment wordt toegepast om een compleet beeld te krijgen van de prestaties van de leerling en om progressie over een bepaalde periode te kunnen meten. Coaches zijn voortdurend in interactie met de leerlingen en kunnen hen stimuleren om hun gedachten te verwoorden. Het beoordelen van cognitieve processen is nodig om diagnostische informatie te verkrijgen over de prestaties en tekort-

komingen in de prestaties in het bijzonder. *Progressierapporten* worden elke week of elke twee weken ingevuld tijdens de opleiding in de preOJT en OJT. De tijdsintervallen zijn verschillend vanwege de diversiteit in de operationele roosters. De coach heeft een dubbelrol als coach en beoordelaar, en meerdere beoordelaars worden ingezet. In de preOJT worden bovendien *simulator tests* gebruikt om de prestaties van de leerling op een objectieve manier te meten, op een vastgesteld moment en in uitgebalanceerde testscenarios zonder tussenkomst van de coach. Pass-fail beslissingen worden genomen aan het eind van elke fase in de preOJT en OJT. Er bestaat geen vaste aftestgrens, maar de beslissingen worden gebaseerd op een uitgebreide kwantitatieve en kwalitatieve analyse van simulator tests en progressierapporten.

Een web-based leerlingvolgsysteem is ontworpen om progressierapporten in te vullen, om de resultaten te kunnen opslaan in een database, en om diverse rapporten te kunnen genereren van de prestaties van leerlingen. Op deze manier kunnen belanghebbenden de voortgang van leerlingen beter volgen vanuit allerlei lokaties, mits zij toegang hebben tot het systeem. De opleidingsresultaten kunnen goed worden gebruikt voor betrouwbaarheids- en validiteitsonderzoek. De eerste versie is gemaakt in het toetssysteem Questionmark Perception. Vanwege technische problemen is echter een nieuw project gestart voor het ontwikkelen van een nieuw leerlingvolgsysteem.

Er is veel energie gestoken in de implementatie van het beoordelingssysteem, omdat dit als erg belangrijk werd beschouwd voor correct gebruik en acceptatie. De nauwe samenwerking met de gebruikers zorgde ervoor dat het systeem vrij gemakkelijk werd geaccepteerd. Bovendien hebben we het systeem zorgvuldig geïntroduceerd door te beginnen met kleine pilots en door het geven van presentaties. We hebben extra aandacht besteed aan beoordeling in de training van coaches en beoordelaars. Tussentijdse evaluaties hebben geleid tot kleine aanpassingen voordat de uiteindelijke versie werd geïmplementeerd.

Evaluatie

De eisen en randvoorwaarden in het PvE hebben geleid tot de evaluatie van het beoordelingssysteem, bestaande uit drie onderdelen. De data zijn van 34 leerlingen in ACC preOJT (188 progressierapporten; 36 simulator tests), en van 27 leerlingen in ACC OJT (407 progressierapporten), verzameld gedurende de periode januari 2001 tot en met december 2006.

Evaluatie van de psychometrische kwaliteit
We hebben de volgende zaken met betrekking tot betrouwbaarheid onderzocht: (1) *interbeoordelaarsbetrouwbaarheid;* (2) *beoordelingsfouten* (halo-effect, toegeeflijkheids- en strengheidsfout, range beperking en centrale tendentiefout); (3) *test betrouwbaarheid* (interne consistentie, split-half betrouwbaarheid). Predictieve validiteit was onderdeel van de evaluatie van leerprocessen (zie volgende paragraaf) omdat hiervoor analyses over de tijd nodig zijn.

De *interbeoordelaarsbetrouwbaarheid* is onderzocht voor simulator tests waarbij twee beoordelaars onafhankelijk van elkaar een beoordeling invullen. Hun ratings moeten uitwisselbaar zijn omdat beoordelingen niet afhankelijk mogen zijn van de specifieke persoon die beoordeelt. We hebben drie typen indicatoren bekeken: vorm, dispersie, en niveau. De resultaten hebben aangetoond dat de interbeoordelaarsbetrouwbaarheid tussen de beoordelaars voldoende hoog is voor het algemene prestatieniveau (gewogen som van ratings op de competenties), maar slechts middelmatig voor de gelijkheid van profielen (vorm, dispersie). Dit houdt in dat sommige beoordelaars lage ratings geven op sommige competenties, terwijl anderen hoge ratings geven op dezelfde competenties en omgekeerd. Beoordelaars vinden het blijkbaar moeilijk om specifieke problemen bij leerlingen aan te wijzen.

Vervolgens is de aanwezigheid van de meest voorkomende beoordelingsfouten onderzocht. Analyses van de *toegeeflijkheids- en strengheidsfout* hebben laten zien dat beoordelaars de neiging hebben tot toegeeflijkheid zoals verwacht: de gemiddelde ratings op de competenties zijn hoger dan het middelste schaalanker. We vonden een paar systematische verschillen in toegeeflijkheid tussen beoordelaars: er bestaan zgn. 'Santa Claus' (hoge gemiddelden) en 'Axeman' (lage gemiddelden) beoordelaars. Bovendien vertonen de meeste beoordelaars *range beperking*. Ze verdelen de ratings niet gelijkmatig over de zespuntsschaal, maar ze geven vaker positieve ratings en vermijden de extreem lage en hoge waarden op de schaal; de *centrale tendentiefout* komt dan ook niet voor. Verschillen tussen beoordelaars in hun verdeling van ratings, uitgedrukt in congruentie-indexen, werden niet gevonden. Tenslotte onderzochten we de aanwezigheid van *halo-effecten*. We vonden hoge intercorrelaties tussen ratings op sommige competenties. We herkenden de indeling van het ATC Performance Model in de resultaten. Specifieke competenties moeten hoge intercorrelaties hebben, omdat ze conceptueel met elkaar samenhangen. Daarom hebben we geconcludeerd dat de aanwezigheid van halo-effecten geen serieus probleem is. In het algemeen werden beoordelingsfouten vaker gevonden in progressierapporten. Dit bevestigt dat simulator tests betrouwbaarder zijn.

Tenslotte hebben we de *interne consistentie* en *split-half betrouwbaarheid* onderzocht. Berekeningen van Cronbach's alpha's en item-total correlaties per competentie hebben laten zien dat de interne consistentie erg hoog is. Slechts een paar criteria moeten verwijderd of aangepast worden. Split-half betrouwbaarheid werd geschat voor simulator tests met behulp van de Spearman-Brown formule. We vonden erg hoge betrouwbaarheidscoëfficiënten. De test betrouwbaarheid is dus voldoende.

De bevindingen leiden tot de conclusie dat de betrouwbaarheid van het beoordelingssysteem voldoende is, hoewel tot op zekere hoogte. Het systeem is goed ontworpen voor wat betreft de indeling in competenties, criteria en standaarden. Echter, de rol van de beoordelaar kan verbeterd worden. We zijn bezig met het organiseren van een training in beoordelen. Hierin wordt aandacht besteed aan het voorkomen van beoordelingsfouten en aan een meer uniform begrip van de competenties.

Evaluatie van leerprocessen

Een goed ontworpen beoordelingssysteem werkt voldoende als de beoordelingsresultaten een juiste afspiegeling zijn van leerprocessen. Patronen en individuele verschillen in leren (bijv. slow starters, leerplateau's) en in prestaties (sterke en zwakke punten) moeten duidelijk onderscheiden worden om te kunnen dienen als basis voor adequate feedback en interventies. Leercurves kunnen worden afgeleid van de beoordelingsresultaten, gebaseerd op een opeenvolging van prestatiemetingen over de tijd. Deze leercurves moeten voldoende representatief zijn voor leerprocessen. In dat geval kan de opleiding maximaal adaptief gemaakt worden aan de behoeftes van de leerling. Pass-fail beslissingen zullen meer valide zijn als ze gebaseerd zijn op de voorspelbaarheid van patronen in leerprocessen.

Leercurves worden normaal gesproken gepresenteerd als groeicurves op basis van herhaalde uitvoering van dezelfde taak op opeenvolgende momenten in de tijd, maar dit is niet mogelijk met ons beoordelingssysteem. De prestaties van de leerling worden beoordeeld ten opzichte van oplopende standaarden die continu worden vertaald naar dezelfde zespuntsschaal. Daarom kunnen de leercurves, geproduceerd door ons beoordelingssysteem, beter gerecalibreerde leercurves worden genoemd. We hebben de leercurves, die afgeleid zijn van de progressierapporten in de preOJT en OJT, vergeleken met prototypische leercurves zoals deze worden verwacht vanuit algemene leertheorie. We hebben drie groepen gedefinieerd: (1) *goede leerlingen* (geslaagd zonder problemen); (2) *matige leerlingen* (geslaagd met problemen); (3) *slechte leerlingen* (gezakt). Drie training managers hebben de leerlingen ingedeeld in de drie groepen, gebaseerd op expert judgment dat diende als extern criterium voor succes in de opleiding. We hebben twee kwantitatieve variabelen gedefinieerd voor de leercurves, onderverdeeld in vier metingen: (1) *prestatie*: gemiddeld prestatieniveau; aantal keer dat de prestaties onvoldoende waren; (2) *progressie:* groei; mate van groei.

We hebben het onderscheid tussen de drie groepen visueel bekeken en ook kwantitatief onderzocht. De resultaten hebben laten zien dat de leercurves van de echte leerlingen voldoende representatief zijn voor leerprocessen, omdat ze voldoende gelijkenis vertonen met de prototypische leercurves. Hoewel het aantal leerlingen nog te laag is om overtuigend bewijs te verkrijgen, hebben we toch verkend wat de voorspelbaarheid was van leercurves over de tijd. De resultaten hebben gesuggereerd dat de leercurves in zekere mate voorspellend zijn voor de volgende fasen. Het is echter nog veel te vroeg om vaste aftestgrenzen op te stellen.

Een goed ontworpen beoordelingssysteem moet niet alleen leerprocessen adequaat weergeven zoals uitgedrukt in de leercurves, maar ook de competentieontwikkeling. Dit is belangrijk voor het ontdekken van mogelijke gebreken hierin. We hebben dezelfde indeling in drie groepen van leerlingen gehanteerd. De bevindingen waren in overeenstemming met de literatuur voor wat betreft de leerbaarheid van competenties. Zoals verwacht vonden we meer significante verschillen tussen individuele leerlingen voor competenties die kritischer en minder leerbaar zijn, zoals mentale beeldvorming en omgaan met werkdruk, dan voor competenties die minder kritisch en meer leerbaar zijn, zoals labelbehandeling en omgaan met apparatuur. De belangrijkste redenen voor aftesten zijn respectievelijk: mentale beeldvorming, omgaan met werkdruk, aandachtsverdeling, en besluitvaardigheid.

We concluderen dat het beoordelingssysteem individuele verschillen in leerprocessen voldoende weergeeft. Meer longitudinaal onderzoek is nodig voor het vinden van bewijzen voor predictieve validiteit.

Gebruikersevaluatie
Meerdere methoden werden gebruikt om het beoordelingssyteem te evalueren met de gebruikers (leerlingen, coaches, training managers). De resultaten hebben aangetoond dat het systeem praktisch bruikbaar is en dat het zeker heeft geleid tot een verbetering van leerprocessen. De middelen, procedures en beoordelingsformulieren zijn duidelijk.

Conclusies en verder onderzoek

De doelstelling om een beoordelingssysteem voor LVNL te ontwerpen dat de opleiding meer efficiënt en effectief maakt is bereikt. Leerprocessen worden beter ondersteund, en pass-fail beslissingen zijn betrouwbaarder en meer valide. Er is voldaan aan het merendeel van eisen en randvoorwaarden van het PvE. We hebben helaas nog geen definitief bewijs dat de slagingspercentages in de opleiding omhoog zijn gegaan.

In het algemeen hebben we meer inzicht verkregen in het ontwerp van beoordelingssystemen voor het aanleren van complexe vaardigheden, gericht op de toepassing in het domein van de luchtverkeersleiding. Een beperking van deze dissertatie is echter dat de resultaten slechts op een klein sample gebaseerd zijn. Sommige zaken vereisen verder onderzoek. Meer onderzoek is nodig naar het ontwerp van beoordelingssystemen voor het aanleren van complexe vaardigheden in simulator opleidingen en werkplekopleidingen, waarbij speciale aandacht geschonken moet worden aan de betrouwbaarheid en validiteit van de ratings van menselijke beoordelaars. Bovendien is meer inzicht nodig in het aanleren van complexe vaardigheden om te bewerkstelligen dat beoordelingen optimale ondersteuning kunnen bieden aan leerprocessen. Onze nieuwe methode om leercurves af te leiden kan een bijdrage leveren aan het onderzoek naar het modelleren van leerprocessen. Daarnaast moet de mate van leerbaarheid van competenties in luchtverkeersleiding en andere domeinen verder onderzocht worden. De relatie met aangeboren cognitieve capaciteiten en persoonlijkheidskenmerken verdient extra aandacht ter verbetering van de predictieve validiteit van de selectie en opleiding. Ook is verder onderzoek vereist naar de relatie tussen adaptief opleiden en mogelijke gebreken in competenties. We hebben veel moeilijkheden ervaren in het adaptief maken van de opleiding aan de behoeftes van leerlingen, terwijl dit een sleutel tot succes kan zijn voor het succesvol afronden van de opleiding. Mogelijke interventies zijn: (dynamische) taakselectie, specifieke coaching, remedial teaching en counseling, alternatieve opleidingstrajecten, verlenging van de opleiding, heropleiden, etcetera. Tenslotte kunnen meer geavanceerde technologie en geautomatiseerde metingen bijdragen tot een verbetering van beoordelingssystemen.